An Overview of Web Archiving

Jinfang Niu
University of South Florida

http://www.dlib.org/dlib/march12/niu/03niu1.html

D-Lib Magazine

March/April 2012
Volume 18, Number ¾

Resumo

Esta visão geral é um estudo dos métodos usados em uma variedade de universidades e bibliotecas e arquivos governamentais internacionais para selecionar, adquirir, descrever e acessar recursos da web para seus arquivos. A criação de um arquivo da Web apresenta muitos desafios, e as bibliotecas e escolas de informação devem garantir que a instrução em métodos e habilidades de arquivamento na Web seja parte de seus currículos, para ajudar os futuros profissionais a enfrentar esses desafios. Na preparação para o desenvolvimento de um curso de arquivamento na web, o autor realizou uma revisão abrangente da literatura. As descobertas são relatadas neste documento, juntamente com as visões do autor sobre alguns dos métodos em uso, tais como conceitos e teorias tradicionais de gerenciamento de arquivos podem ser aplicados à organização e à descrição de recursos da Web arquivados.

Introdução

O arquivamento na Web é o processo de reunir dados que foram registrados na World Wide Web, armazenando-os, garantindo que os dados sejam preservados em um arquivo morto e disponibilizando os dados coletados para pesquisas futuras. O Internet Archive e várias bibliotecas nacionais iniciaram práticas de arquivamento na web em 1996. O International Web Archiving Workshop (IWAW), iniciado em 2001, forneceu uma plataforma para compartilhar experiências e trocar ideias. A fundação posterior do International Internet Preservation Consortium (IIPC), em 2003, facilitou muito a colaboração internacional no desenvolvimento de padrões e ferramentas de código aberto para a criação de arquivos da web. Esses desenvolvimentos e a crescente parcela da cultura humana criada e registrada na web se combinam para tornar inevitável que mais e mais bibliotecas e arquivos tenham que enfrentar os desafios do arquivamento na web.

As escolas de informação e biblioteca precisam preparar os alunos para esses desafios. Uma pesquisa dos catálogos de cursos das 32 melhores escolas de informação e biblioteca dos EUA, no outono de 2010, descobriu que apenas uma escola, a Universidade de Michigan, ofereceu um curso semestral sobre arquivamento na web. A Universidade de Indiana abordou o arquivamento na Web como um tópico em seu curso "Análise de

conteúdo para a Web". A UCLA cobriu o arquivamento na Web como um tópico em seu curso de 'Gerenciamento de Registros Digitais' (US News e World Report Weekly, 2009). Embora muitas escolas, por exemplo, a Universidade de Illinois, ofereçam cursos sobre preservação digital, curadoria digital e o impacto da web 2.0 na teoria e na prática de arquivamento, não se sabe até que ponto qualquer deles aborda problemas de arquivamento na web. O autor acredita que o arquivamento na web requer conhecimento e habilidades únicas suficientes para exigir um curso separado. Na preparação para o desenvolvimento de um curso de arquivamento na web, o autor realizou uma revisão abrangente da literatura e avaliou as funcionalidades de vários arquivos da web proeminentes. Este artigo, e um segundo artigo também publicado na revista D-Lib, "Functionalities of Web Archives", resultou da pesquisa de preparação do curso.

Como o gerenciamento de muitos outros tipos de recursos de informações, o fluxo de trabalho do arquivamento na web inclui avaliação e seleção, aquisição, organização e armazenamento, descrição e acesso. Esse fluxo de trabalho é o núcleo do arquivamento na web. As seções a seguir revisam os métodos usados em cada etapa do fluxo de trabalho e apresentam os pontos de vista do autor sobre alguns desses problemas. Embora a preservação digital seja definitivamente uma etapa importante no processo geral de arquivamento na web, ela não é exclusiva do arquivamento na web. A preservação dos recursos da web não é diferente da preservação de outros recursos digitais. Ele pode ser coberto em um curso de biblioteca digital ou também em um curso de gerenciamento de registros eletrônicos. Portanto, esta revisão não cobre a preservação digital.

Avaliação e Seleção

O termo avaliação é usado na comunidade de arquivos para se referir ao processo de avaliação do valor dos registros e para decidir se e por quanto tempo os registros devem ser preservados. É essencialmente um processo de seleção. Neste artigo, é usado como sinônimo de seleção. Todos os arquivos da web selecionam recursos da web para preservar com base em um ou mais critérios. Embora o Internet Archive tente arquivar toda a Web, ele apenas captura páginas da Web na superfície da Web (Lecher, 2006). As páginas da Web mais abaixo na hierarquia de sites geralmente não são coletadas pelo Arquivo da Internet.

Os esforços de arquivamento da Web existentes usam os seguintes critérios de seleção para determinar o que preservar: domínio (como .gov ou .edu), tópico ou evento, tipo de mídia e gênero. Muitos países europeus arquivam a web em seu domínio de país. A biblioteca do Centro de Voo Espacial Goddard da NASA (GSFC) captura páginas no domínio Goddard (Senserini et al., 2004). A Biblioteca do Congresso criou várias coleções na web baseadas em eventos, como os arquivos da web de 11 de setembro de 2001, os arquivos da web para a eleição e os arquivos da web da Guerra do Iraque 2003 (Biblioteca do Congresso, 2011). A seleção baseada no tipo de mídia inclui ou exclui determinados tipos de mídia. A biblioteca Goddard, por exemplo, evita o rastreamento de grandes arquivos de vídeo e produtos de software (Senserini et al., 2004). O projeto de arquivamento da web conduzido por Chirag Shah e Gary Marchionini (2007), por outro lado, focou na preservação de vídeos eleitorais no Youtube. Alguns arquivos da web são

selecionados com base em gêneros como blogs, jornais, mundos virtuais, etc. A Biblioteca Nacional da França criou uma coleção eletrônica de e-diários (Lasfargues et al., 2008). O Arquivo da Internet possui um arquivo de software e um arquivo de vídeos de videogame (Internet Archive, 2001a; Internet Archive, 2001b). O projeto Preserving Virtual Worlds realiza pesquisas especificamente sobre o arquivamento de mundos virtuais on-line (Preserving Virtual Worlds, 2008). Antonescu, et al. (2009) apontou duas abordagens diferentes para preservar os mundos virtuais online. Uma abordagem preserva a infraestrutura técnica - os objetos e os avatares existentes nos mundos virtuais - enquanto a outra abordagem preserva a interação e as experiências de vida dos avatares nos mundos virtuais. Winget e Murray realizaram pesquisas para preservar os registros e artefatos criados durante o processo de desenvolvimento de videogames (Winget e Murray, 2008).

Teoricamente, a seleção baseada em critérios objetivos pode ser facilmente automatizada. Em um nível técnico, é fácil para o software decidir o tipo de mídia (áudio, vídeo ou texto) e o domínio (por exemplo, .gov ou .au) dos recursos da web. Da mesma forma, não deve ser muito difícil diferenciar entre gêneros como diários on-line ou blogs, ou perceber as diferenças entre as postagens do blog e os comentários. É possível identificar conteúdo da web de alta qualidade ou popular com base no número de links e visitantes recebidos, número de espectadores de vídeos on-line e classificações de usuários. A Biblioteca Nacional da República Tcheca automatizou a identificação da Web tcheca fora do domínio nacional, que inclui sites tchecos que não estão no domínio .cz, mas em domínios .net, .com, .org ou .edu (Vlcek, 2008). Um WebAnalyzer foi criado e integrado ao rastreador. Durante o rastreamento, o WebAnalyzer analisa páginas da web e procura por algumas propriedades pré-definidas que caracterizam a web tcheca. Toda vez que uma propriedade predefinida é encontrada, uma certa quantidade de pontos é adicionada ao URL. Quando um certo limite é atingido, a página da web é considerada parte da web tcheca e será arquivada.

Seleção baseada em tópico ou evento, no entanto, precisa de julgamento humano. A seleção manual por profissionais da informação é demorada e cara e, portanto, é usada apenas em arquivos da Web de pequena escala. Para reduzir o custo da seleção manual, alguns arquivos da Web aceitam URLs recomendados pelo usuário, usam registros existentes de URLs ou envolvem especialistas em assunto para ajudar na seleção de recursos da Web para arquivamento. A Preservação e Acesso aos Recursos Documentários em Rede da Austrália (PANDORA) e os arquivos da Web da Biblioteca Nacional da Universidade de Taiwan aceitam sites recomendados pelo usuário (Biblioteca Nacional da Austrália, 2008; Chen et al., 2008). O Arquivo Digital para Estudos Chineses (DACHS) convidou acadêmicos especialistas em estudos chineses para recomendar sites relacionados (Lecher, 2006). O projeto de arquivamento na web do governo do Reino Unido seleciona sites usando um registro de todos os sites do governo central do Reino Unido; As URLs no registro são enviadas e mantidas pelos gerentes do site (Spencer et al., 2009).

Outra maneira de acelerar a seleção manual é usar a teoria de macroavaliação no campo de gerenciamento de arquivos. Conforme explicado no Modelo do Arizona para curadoria de publicações da web do governo, a avaliação macro envolve avaliação e seleção de recursos da web com base em agregados de páginas da web em vez de páginas da web individuais

(Pearce-Moses e Kaczmarek, 2005). A avaliação de agregados reduz o tamanho do problema e torna o processo de avaliação mais eficiente. Os agregados podem ser decididos em diferentes níveis. A Administração Nacional de Arquivos e Registros dos EUA (NARA) utilizou várias unidades de análise em sua orientação para que agências governamentais conduzissem análises de risco para registros da Web: grupo de sites, um site inteiro, um site menos uma ou duas partes que exibissem características substancialmente diferentes. e clusters de páginas web (NARA, 2005). Essas várias unidades de análise também podem ser aplicadas na seleção de arquivamento na web. Por exemplo, bibliotecários ou arquivistas podem avaliar o valor de um site inteiro em vez de páginas individuais da web para decidir se o site deve ser arquivado.

Os critérios de seleção, como domínio ou tipo de mídia, podem ser associados a uma seleção baseada em valor ou a um método de amostragem representativo. O arquivo da web da Universidade Nacional de Taiwan reúne recursos da web que são valiosos de pontos de vista históricos, culturais, sociais, educacionais ou acadêmicos (Chen et al., 2008). A filtragem de spam também é um tipo de método de seleção baseado em valor. A amostragem representativa, por outro lado, evita a subjetividade e o viés na avaliação baseada em valor e tenta criar uma imagem representativa do que deve ser preservado. Lyle (2004) aplicou a estratégia de amostragem a recursos da Web que foram baixados por rastreadores como uma forma de reduzir a quantidade de recursos da Web a serem arquivados. A Biblioteca Nacional da França usou a estratégia de amostragem para decidir a lista de sementes e os critérios de filtragem antes de rastrear; A Biblioteca Nacional acredita que as coleções devem "espelhar a sociedade e a cultura francesas em toda a sua diversidade, independentemente do valor científico ou da popularidade das publicações" (Lasfargues et al., 2008). Devido a esta crença, "o arquivo web inclui o 'melhor' (literatura, publicação científica) assim como o 'pior' (de propagandas à pornografia). Pequenas, médias e grandes têm a mesma chance de serem coletadas" (Lasfargues et al., 2008).

Aquisição

Dependendo da escala do arquivo da Web, da relação entre o arquivo da Web e os proprietários do site e a natureza do conteúdo da Web arquivado, diferentes métodos de aquisição podem ser usados para obter o conteúdo da Web. Bibliotecas e arquivos têm uma longa tradição de aceitar transferências de agências governamentais, doadores e depósitos legais de editores. Este método ainda se aplica ao arquivamento da web. Por exemplo, o NARA pediu a todos os departamentos federais que entregassem um instantâneo de seu site para o NARA até o final do mandato do presidente Clinton (Bellardo, 2001). Adrian Brown (2006) apontou que os sites dinâmicos baseados em banco de dados não são adequados para transferência direta, porque os bancos de dados são geralmente proprietários e difíceis de preservar a longo prazo. Uma abordagem mais fácil é converter dados de bancos de dados proprietários em um formato padrão aberto, como XML, usando uma ferramenta como o DeepArc.

Um método de aquisição exclusivo para arquivamento na web está sendo rastreado. Esse método depende de rastreadores para coletar conteúdo de servidores da web. Os rastreadores usam uma lista de origens para iniciar o download do conteúdo da Web e

seguem os hiperlinks para descobrir e baixar o conteúdo da Web adicional. As decisões de seleção são a base para compilar uma lista de sementes e configurar os parâmetros do rastreador. Por exemplo, a Biblioteca Nacional da França decidiu rastrear todos os sites nos domínios de primeiro nível .fr e .re e qualquer outro domínio que tenha sido redirecionado de um domínio .fr ou .re (Lasfargues et al., 2008). Essa decisão de seleção é configurada nos rastreadores como um filtro. Apenas os links que passam pelo filtro serão arquivados. O rastreamento está substituindo o depósito na aquisição de publicações da Web em algumas bibliotecas e arquivos. Por exemplo, a Biblioteca Estadual do Arizona mudou para rastreamento de espera de submissões de agências do governo estadual (Pearce-Moses e Kaczmarek, 2005). Alguns recursos da Web precisam ser capturados manualmente devido às limitações do rastreador. Por exemplo, alguns rastreadores não podem coletar arquivos GIS, conteúdo dinâmico da web ou fluxo de mídia. O NARA fornece um guia para métodos de captura apropriados de formatos de registro de conteúdo da web específicos que não podem ser capturados por rastreadores (NARA, 2004).

O rastreamento repetido de páginas não atualizadas gera duplicatas no arquivo da Web, o que desperdiçará recursos para gerenciamento, armazenamento e preservação. Felizmente, os rastreadores inteligentes atuais, como a versão atual do Heritrix, têm a funcionalidade de reduzir as duplicatas no download e no armazenamento de recursos da web. O rastreamento repetido de sites grandes e atualizados com freqüência causa incoerência temporal. Pode levar vários dias ou até mais para rastrear um site grande, durante o qual os sites estão passando por alterações. Suponha que haja duas páginas da Web (p1 e p2) em um site. O rastreador fez o download de p1 no horário t1. Quando o rastreador atinge p2, p2 e p1 foram atualizados para p2-a e p1-a, respectivamente. Nesse cenário, o site original inclui p1 e p2, o site atualizado contém p1-ae p2-a, mas o site arquivado inclui p1 e p2-a. Em outras palavras, o rastreador coletou um site que nunca existiu. Pesquisas estão sendo conduzidas para reduzir a incoerência temporal em arquivos da web (Spaniol et al., 2008).

Ao adquirir recursos da Web, a decisão de solicitar permissão de proprietários de direitos autorais depende do ambiente legal do arquivo da Web, da escala do arquivo da Web e da natureza do conteúdo arquivado e da organização de arquivamento. Em um país onde o depósito legal cobre recursos da web, como a Nova Zelândia, a biblioteca de depósito legal não precisa buscar permissão para arquivar publicações da Web produzidas naquele país. Os arquivos do governo que têm o mandato legal para preservar registros públicos, como o NARA e o Arquivo Nacional do Reino Unido, também não precisam pedir permissão aos produtores de discos. No mesmo ambiente legal, é mais provável que a busca de permissão seja realizada para arquivamento da Web em pequena escala, em vez de em grande escala, porque é mais fácil solicitar a permissão de um número relativamente pequeno de proprietários de direitos autorais. Arquivos da Web de grande escala, como o Arquivo da Internet, tendem a usar o mecanismo de desativação (obedeça à exclusão de robôs e permita a remoção de solicitações). Hal Varian (2006) argumentou que o mecanismo de desativação do Projeto de Biblioteca do Google é uma escolha sensata, porque o custo de transação do modelo opt-in, no qual a permissão é solicitada, é alto demais para ser bem-sucedido. Esse argumento também é válido para o arquivamento na Web, talvez ainda mais porque a maioria dos arquivos da Web não se beneficia financeiramente com o

arquivamento de conteúdo da Web e é difícil identificar os proprietários dos direitos autorais de conteúdo da Web postados por usuários anônimos.

A escala do esforço de arquivamento na Web também afeta a decisão de obedecer à exclusão de robôs. De acordo com uma lei de direitos autorais de 2006 da França, a Biblioteca Nacional da França pode ignorar a exclusão de robôs enquanto rastreia a web francesa (Lasfargues et al., 2008). Na prática, a Biblioteca Nacional da França geralmente não obedece à exclusão de robôs ao executar rastreamentos focados em pequena escala, mas obedece à exclusão de robôs em rastreamentos amplos porque é mais fácil gerenciar as consequências (como protestos de proprietários de sites e rastreadores de rastreadores associados) de ignorar a exclusão de robôs no arquivamento da Web de pequena escala do que no arquivamento da Web em grande escala (Lasfargues et al., 2008). A natureza do conteúdo arquivado também afeta a decisão de solicitar permissão ou não. A Biblioteca do Congresso procura permissão para arquivar blogs e sites de organizações de notícias, mas apenas notifica a maioria dos outros tipos de sites que a biblioteca está arquivando em seus sites (Grotke e Jones, 2010).

Organização e Armazenamento

Os arquivos da Web precisam preservar a autenticidade e a integridade do conteúdo da Web arquivado. Os requisitos de autenticidade e integridade variam de acordo com o objetivo da coleta. Em alguns cenários, preservar apenas o conteúdo intelectual é suficiente. Em outros cenários, como na preservação de evidências legais, a estrutura e o contexto dos recursos também podem precisar ser preservados. Na teoria tradicional de gerenciamento de arquivos, o contexto dos registros de arquivos inclui a proveniência e a ordem original. A proveniência inclui informações sobre a origem dos registros, como os produtores de registros, as transações que fazem com que os registros sejam produzidos e a cadeia de custódia. A ordem original é a ordem na qual os produtores de registros ou os gerentes de registros organizaram originalmente os registros para demonstrar as relações entre os registros. Embora muitos arquivos da Web preservem o conteúdo da Web como recursos de informação, e não como evidência, o conceito de proveniência ainda se aplica. Para recursos da web arquivados, a proveniência inclui a URL de um site, os produtores de conteúdo e a transação comercial ou a finalidade que causou a produção dos recursos da web. O URL é um metadado externo associado ao recurso da web. Outras informações sobre proveniência são frequentemente incorporadas no conteúdo do recurso da web.

Para recursos da web, o conceito de ordem original pode ser combinado com o conceito de estrutura definido na teoria tradicional de gerenciamento de arquivos. A ordem original é essencialmente a estrutura externa do objeto da web arquivado. A estrutura definida na teoria tradicional de gerenciamento de arquivos é essencialmente a estrutura interna do objeto da web arquivado. Por exemplo, para um site arquivado, sua estrutura externa mostra como este site é organizado em relação a outros sites, o que também pode ser considerado como a ordem original do site. Links de entrada que vêm de fora do site e links de saída deste site para outros sites são parte desta estrutura externa e, portanto, a ordem original deste site. A estrutura hierárquica interna do site mostra como os componentes e subcomponentes deste site são organizados em relação uns aos outros, o que pode ser

considerado como a estrutura definida na teoria tradicional de gerenciamento de arquivos. Essa estrutura interna é definida pelos hiperlinks no site. Para um objeto arquivado de nível inferior, como uma página da Web, a estrutura externa mostra como essa página da Web é organizada em relação a outras páginas da Web. Os links de saída desta página da Web e os links recebidos de fora desta página definem a estrutura externa e a ordem original desta página da Web. A estrutura interna mostra como os componentes internos desta página web, por exemplo, como o conteúdo textual, imagens, áudio, vídeos. etc, estão dispostos. Em colheitas repetidas, o contexto histórico que mostra como o conteúdo da Web evoluiu também existe. Inclui as versões mais antigas e mais recentes das páginas da web.

De acordo com Masanes (2006), os arquivos da web atuais usam principalmente três abordagens para organizar e armazenar conteúdo da web arquivado: sistemas de arquivos locais, arquivos baseados na web e arquivos não baseados na web. Todas as três abordagens preservam o conteúdo intelectual das páginas da Web, mas variam no grau de preservação do contexto e da estrutura.

Em um arquivo da Web que usa um sistema de arquivos local, o navegador pode navegar no sistema de arquivos da mesma forma que navega na Web (Masanes, 2006). Tanto a estrutura hierárquica interna dos sites quanto os relacionamentos de links entre os diferentes sites são preservados, exceto os links não arquivados que estão fora do escopo do arquivo da web. No entanto, duas transformações contextuais precisam ser feitas para permitir que recursos da Web se encaixem em sistemas de arquivos. Primeiro, a nomenclatura dos URIs precisa ser modificada para estar em conformidade com as regras dos sistemas de arquivos locais. Segundo, os links absolutos precisam ser transformados em links relativos para permitir a navegação dentro do sistema de arquivos, caso contrário, os links absolutos apontarão para páginas da Web ativas em vez de conteúdo arquivado.

Em um arquivamento baseado na Web, as páginas da Web e os metadados associados são agrupados e armazenados em arquivos de contêiner e os links e URIs originais são preservados. Embora os links também precisem ser redirecionados ou transformados para apontar para o arquivo morto, e não para a Web ativa, o redirecionamento ou transformação do link ocorre apenas quando os usuários acessam esses links, em vez de precisarem ser gravados no arquivo morto. Esta segunda abordagem preserva a autenticidade ao maior grau.

A abordagem de arquivamento não baseado na Web extrai documentos da Web do contexto de hipertexto e reorganiza-os em um modo de acesso baseado em catálogo ou os transforma em arquivos PDF. Essa abordagem preserva a autenticidade e a integridade ao menor grau.

Descrição e Metadados

A abordagem de geração de metadados e a riqueza de metadados gerados variam de acordo com a escala do arquivo da Web e os recursos disponíveis na organização de arquivamento. Arquivos web muito grandes geralmente dependem da geração automática de metadados. Algumas informações de metadados, como o timestamp gerado quando o

recurso da Web foi coletado, o código de status (por exemplo, 404 para não encontrado ou 303 para redirecionamento), o tamanho em bytes, o URI ou o tipo MIME (por exemplo, text / html), pode ser criado ou capturado por rastreadores. As informações de metadados também podem ser extraídas das meta tags de páginas HTML, embora algumas metatags não sejam precisas devido à Otimização do Mecanismo de Pesquisa. O projeto do Grego da Web extrai automaticamente palavras-chave de páginas da Web e texto âncora e, em seguida, usa as palavras-chave para classificar páginas da Web em clusters (Lampos et al., 2004).

Arquivos da Web em pequena escala podem criar metadados manualmente. O arquivo de literatura de campanha on-line da Universidade da Califórnia em Los Angeles usa o padrão de metadados Dublin Core, cabeçalhos de assunto da Biblioteca do Congresso e listas de autoridade definidas localmente. Seus metadados administrativos são derivados das notas detalhadas criadas pela equipe durante o processo de captura e revisão (Gray e Martin, 2007). O arquivo digital dos arquivos da web de Estudos Chineses convidou os acadêmicos a contribuírem com alguns metadados descritivos (Lecher, 2006). Os Arquivos da Web da Universidade Nacional de Taiwan criaram um esquema de classificação hierárquica de três níveis e regras de catalogação especialmente para o conteúdo da web (Chen et al., 2008). Os metadados também podem ser criados por meio de marcação, comentário ou classificação do usuário. A Biblioteca do Congresso gera automaticamente registros do Esquema de Descrição de Objeto de Metadados (MODS) com base em metadados criados por nominadores de URL e, em seguida, aprimora os registros por catalogadores (Grotke e Jones, 2010).

Coleções de arquivos da Web têm uma estrutura hierárquica de vários níveis. Uma coleção de arquivos da Web pode incluir várias sessões de rastreamento. Em cada sessão de rastreamento, vários sites são rastreados. Cada site inclui muitas páginas da web. Cada página da Web pode ser composta de muitos arquivos, como um arquivo de texto, um arquivo de imagem e um arquivo de vídeo. Essa estrutura hierárquica corresponde à estrutura hierárquica de uma coleção de archive. Os métodos de descrição multinível usados para arquivos podem ser aplicados a sites arquivados. A comunidade de arquivos usa uma abordagem de cima para baixo: os metadados são criados para os níveis mais altos primeiro; depois, se os recursos estiverem disponíveis, os metadados para o nível inferior serão criados; metadados criados para níveis mais altos podem ser herdados por níveis mais baixos; os metadados são raramente criados para objetos no nível do item. Essa abordagem de cima para baixo e o mecanismo de herança de metadados também podem ser aplicados a arquivos da web. Além disso, alguns metadados para os objetos de nível de item, como formato de arquivo, tamanho em bytes e data de modificação, podem ser extraídos automaticamente.

No caso em que um arquivo da web decide usar uma abordagem bibliográfica e criar apenas uma descrição de nível único, ele deve escolher a unidade de descrição com base na escala dos arquivos da web e dos recursos disponíveis. Uma unidade de descrição em um nível superior, como um site inteiro, significa uma descrição menos detalhada e menos registros de metadados serão criados. O arquivo da Web da Biblioteca do Congresso e da Universidade de Harvard cria um registro MARC para uma coleção de arquivos da Web que

inclui muitos sites. Este registro MARC é pesquisável através do catálogo da biblioteca. Uma unidade de descrição em um nível inferior, como os resultados no nível da página em uma descrição mais detalhada e mais registros de metadados, será criada. Além dos registros MARC de uma coleção de arquivos da Web, o arquivo da Web da Biblioteca do Congresso também cria registros de MODS para sites (Biblioteca da Web do Congresso, 2011). Esses registros de MODS são pesquisáveis no arquivo da Web, mas não podem ser acessados pelo catálogo da biblioteca. A PANDORA também escolhe um site e uma parte de um site como unidades de descrição (Hallgrimsson, 2006).

Acessar e usar

A acessibilidade dos arquivos da web depende do ambiente legal do país no qual o arquivo está hospedado. A legislação de depósito legal da Nova Zelândia permite que a Biblioteca Nacional da Nova Zelândia preserve quaisquer páginas disponíveis publicamente de um site da Nova Zelândia e forneça acesso à cópia arquivada do site (Biblioteca Nacional da Nova Zelândia, 2010). Nos EUA, a Biblioteca do Congresso torna os registros bibliográficos de todos os sites arquivados publicamente acessíveis e só pode fornecer acesso público a páginas da Web cujos produtores deram permissão (Grotke e Jones, 2010). Muitos arquivos da web são arquivos obscuros ou apenas acessíveis no local, como os arquivos da web da Biblioteca Nacional da França e do Institut National de l'Audiovisuel (INA) da França, o arquivo web finlandês, Netarchive.dk, Web Archive Norway, o Webarchive da Eslovénia, Web Archive Suíça e Web Archive Áustria (International Internet Preservation Consortium, 2011). Alguns arquivos da web acessíveis ao público oferecem funcionalidade reduzida e acesso atrasado para evitar a concorrência com os proprietários de sites (Masanes, 2006). Por exemplo, há um atraso de pelo menos três meses entre o momento em que um site é coletado e quando ele será exibido no WAX (Harvard University Library, 2009). No caso da IA Wayback Machine, o atraso é de 6 a 12 meses (Archive-it, 2011).

O recurso de pesquisa de diferentes arquivos da web depende da riqueza de metadados e das ferramentas de pesquisa e indexação usadas. O arquivo da web da Biblioteca do Congresso e o arquivo da web da Nova Zelândia dão suporte à pesquisa por meio de pontos de acesso controlados por autoridades. Isso foi possível graças ao fato de que esses dois arquivos da web usavam cabeçalhos de assunto em seus registros de metadados para sites arquivados. Os arquivos da Web baseados no Wayback Machine, por outro lado, são pesquisáveis apenas por URL, enquanto os arquivos da Web baseados no mecanismo de pesquisa NutchWax também podem oferecer suporte à pesquisa de texto completo. Algumas interfaces de acesso avançadas foram criadas. O arquivo da web do Reino Unido criou duas interfaces de visualização para seus arquivos da web com base no conteúdo de mineração, nuvens de tags e uma parede 3D (UK Web Archive, 2011). Jatowt et al. (2008) também experimentaram vários métodos avançados para exibir as versões históricas das páginas da web; eles criaram uma apresentação de slides e um gráfico bidimensional para mostrar como o conteúdo de uma URL evoluiu ao longo do tempo.

Conclusão e o próximo passo

Os arquivos da web existentes demonstram uma variedade de métodos e abordagens para selecionar, adquirir, organizar, armazenar, descrever e fornecer acesso. Essa variação é causada por fatores externos, como o ambiente jurídico e os relacionamentos entre os produtores de recursos da Web e o arquivo da Web, além de fatores internos, como a natureza do conteúdo da Web arquivado, a natureza da organização de arquivamento e a escala de o arquivo da Web e a capacidade técnica e financeira da organização de arquivamento.

Esta visão geral é baseada em uma revisão abrangente da literatura que explica como o arquivamento da web está sendo feito. No entanto, nenhuma literatura aborda diretamente os conhecimentos e habilidades requeridos pelos profissionais da área que realizam a rotina diária de seleção, aquisição e catalogação de arquivos da web. O autor está planejando um projeto de pesquisa para preencher essa lacuna, para o qual bibliotecários e arquivistas que realizam essas tarefas serão entrevistados. Os pontos de vista dos profissionais fornecerão informações adicionais valiosas para o design do curso de arquivamento na Web que está sendo desenvolvido a partir dos resultados dessa pesquisa bibliográfica e de uma avaliação das funcionalidades do arquivo da web.

References

[1] Antonescu, M., Guttenbrunner, M., & Rauber, A. (2009). Documenting a Virtual World — A case study in preserving scenes from SecondLife. Proceedings from IWAW '09: *9th International Web Archiving Workshop*, Corfu (pp. 5-10).

[2] Archive-it. (2011). FAQ. *Archive-it*.

[3] Bellardo, L. J. (2001). *Memorandum to Chief Information Officers: Snapshot of public web sites*.

[4] Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. London: Facet Publishing.

[5] Chen K. H., Chen, Y. L., & Ting, P. F. (2008). Developing National Taiwan University Web Archiving System. Proceedings from IWWA '08: *8th International Workshop for Web Archiving*, Denmark (pp. 1-8).

[6] Gray, G., & Martin, S. (2007). The UCLA Online Campaign Literature Archive: A case study. Proceedings from IWAW '07: *7th International Web Archiving Workshop, Vancouver* (pp 1-5).

[7] Grotke, A., & Jones, G. (2010). DigiBoard: A tool to streamline complex web archiving activities at the Library of Congress. Proceedings from IWAW '10: *10th International Web Archiving Workshop*, Vienna.

[8] Hallgrimsson, T. 2006. Access and finding aids. In *Web Archiving* (pp.131-152). Berlin: Springer-Verlag.

[9] Harvard University Library. (2009). WAX Public Interface Help. Harvard University.

[10] Internet Archive. (2001). Software Archive. *Internet Archive*.

[11] Internet Archive. (2001). Videogame Video Archive. *Internet Archive*.

[12] International Internet Preservation Consortium. (2011). Member Archives. *International Internet Preservation Consortium*.

[13] International Internet Preservation Consortium Access Working Group. (2006). Use cases for access to Internet Archives. *International Internet Preservation Consortium*.

[14] Jatowt, A., Kawai, Y., & Tanaka, K. (2008). Using page histories for improving browsing the Web. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[15] Lampos, C., Eirinaki, M., Jevtuchova, D., & Vazirgiannis, M. (2004). Archiving the Greek Web.

[16] Lasfargues, F., Oury, C., & Wendland, B. (2008). Legal deposit of the French Web: Harvesting strategies for a national domain. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[17] Lecher, Hanno E. (2006). Small scale academic web archiving: DACHS. In *Web Archiving* (pp. 213-226). Berlin: Springer-Verlag.

[18] Library of Congress Web Archives. (2011). Minerva. *Library of Congress*.

[19] Library of Congress Web Archives. (2011). Technical Information. *Library of Congress*.

[20] Lyle, J. (2004). Sampling the Umich.edu Domain. Proceedings from IWAW '04: *4th International Web Archiving Workshop*, Bath, UK.

[21] Masanes, J. (Ed.). (2006). *Web Archiving*. Berlin: Springer-Verlag.

[22] NARA. (2004). Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: Web Content Records. *NARA*.

[23] NARA. (2005). NARA Guidance on Managing Web Records. *NARA*.

[24] National Library of Australia. (2008). Services to researchers. *National Library of Australia*.

[25] National Library of New Zealand. (2010). Frequently Asked Questions About Archives Websites. *New Zealand Web Archive*.

[26] Niu, J. (2012). Functionalities of web archives. *D-Lib Magazine*, Vol.18, No. 3/4. http://dx.doi.org/10.1045/march2012-niu2

[27] Pearce-Moses, R., & Kaczmarek, J. (2005). An Arizona model for preservation and access of Web documents.

[28] Preserving Virtual Worlds. (2008). *Preserving Virtual Worlds*.

[29] Senserini, A., Allen, R. B., Hodge, G., Anderson, N., & Smith, D. Jr. (2004). Archiving and accessing web pages: The Goddard library web capture project. *D-Lib Magazine*, 10 (11). http://dx.doi.org/10.1045/november2004-hodge

[30] Shah, C., & Marchionini, G. (2007). Preserving 2008 US Presidential Election Videos. Proceedings from IWAW '07: *7th International Web Archiving Workshop*, Vancouver, Canada.

[31] Spaniol, M., Denev, D., Mazeika, A., & Weikum, G. (2008). Catch me if you can: temporal coherence of Web archives. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark.

[32] Spencer, A., O'Reilly, B., & Vasile, G. (2009). Past and present: Using the UK Government Web Archive to bridge the continuity gap. Proceedings from IWAW '09: *9th International Web Archiving Workshop*, Corfu. (pp. 38-43).

[33] UK Web Archive. (2011). Visualization. *UK Web Archive*.

[34] US News and World Report Weekly. (2009). Best Graduate Schools: Library and Information Science. *US News and World Report Weekly*.

[35] Varian, H. R. (2006). *The Google Library project*.

[36] Vlcek, I. (2008). Identification and archiving of the Czech Web outside the National Domain. Proceedings from IWAW '08: *8th International Workshop for Web Archiving*, Denmark. Retrieved from http://iwaw.europarchive.org/08/IWAW2008-Vlcek.pdf

[37] Winget , M.A. & Murray, C. (2008). *Collecting and preserving videogames and their related materials: A review of current practice, game-related archives and research projects*.